Concept Chain Graphs for Text Mining

WHITE PAPER

Submitted to:

Ernest Lucier
Ernest.lucier@faa.gov
Advisor on High Confidence Systems
FAA/AIO-4
800 Independence Avenue SW
Washington, DC 20591
202 385-8157

Submitted by:

PI: Rohini K. Srihari
Assoc. Prof. Dept. of Computer Science & Eng
State University of New York at Buffalo
rohini@cedar.buffalo.edu
Ph: (716)-645-6164 x 102; Fax: (716)-645-6176

Project Overview:

It is often the case that a document collection reveals interesting information other than what is explicitly stated. Since many authors working independently and at various times generate these documents, interesting links that connect facts, assertions or hypotheses may be missed. In some cases the hidden information may be an unapparent consequence of multiple sources and authors that may pose issues with respect to sensitive information. We refer to this special case of text mining *as unapparent information revelation (UIR)*. The goal of information analysts is to sift through these extensive document collections and find such links and detect threat scenarios. Currently they perform this task with limited assistance from tools such as search engines. What is required is a set of automated tools that will expose such links, or at least generate plausible concept chains. The UIR problem manifests itself in different scenarios including complex questions, answers, and summarizations.

One specific UIR application that is being tackled concerns sensitive information on web sites. With an increasing number of documents being generated by different individuals and departments in organizations, there is a potential of the release of information unintended for public consumption. This may be inconsistent with the overall objectives and operation of the organization. We refer to this situation as *unintended information revelation*. For example, there is a large volume of information on the FAA web site, www.faa.gov. By following the links, portions of undesirable scenarios can be detected and connected (e.g., fuel leaks can cause onboard explosions and fueling tankers on icy runways are associated with fuel leaks). These links identify vulnerabilities that could be exploited.

We are developing a concept chain graph (CCG) process that has the robustness and scalability of information retrieval frameworks (e.g., the vector space model), in conjunction with richer representation and reasoning frameworks (e.g., Bayesian belief networks). The solution being pursued takes the view that the concepts and associations (i.e., scenarios) in a particular domain can be represented as a probabilistic network with the nodes representing concepts and edges between them representing generic associations. The CCG reflects both concepts and associations derived from the document collection, as well as concepts and associations from the domain ontology.

Anticipated Benefits to the FAA

The goal of this project is to develop a system to detect various types of UIR scenarios. The scenarios of interest can be either *predefined* or discovered in a *bottom-up* manner. The former category corresponds to specific scenarios of interest to the FAA: e.g., whether it is possible to deduce that some airport may be vulnerable due to less funding for security equipment and measures. The bottom-up scenarios are based on graph mining and detecting any *anomalies* with respect to the unusual degree of influence of some concept, or set of concepts. In both cases, these scenarios are constructed from links that span multiple documents. Specifically, the goal is to develop an application that is useful in website information analyses (e.g., exposing and discovering exploitable vulnerabilities). This entails the following:

- A set of tools (UIR toolkit) that allows users to index large document collections (e.g., websites) and construct CCGs
- A web-enabled UIR prototype application with Graphical User Interface (GUI) focused on text mining for homeland security allowing users to:
    - Interactively visualize, query, and browse the CCG

- Search for concept chains (i.e., scenarios) using an ontology of concepts. Responses are lists of text snippets as part of an *evidence trail*.
- Retrieve documents/sub-graphs corresponding to either model-driven scenarios or bottom-up generation of possible scenarios. Bottom-up generation of possible scenarios involves creating ontologies automatically.
- Conduct scenario-mining experiments designed to uncover anomalies, trends, correlations, etc.

This will permit the FAA to use the UIR system for various applications including:

- Web site traffic monitoring for anomalous activity that could uncover a possible threat
- Examination of publicly disseminated information for UIR (e.g., system vulnerabilities)
- Text mining on National Transportation Safety Board (NTSB) accident data to discover new correlations
- Processing of FAA documents central to congressional testimony. This application can provide automatic summarization and analysis of content.

Current Status

We initially focused on UIR scenarios that are more directly relevant to homeland security. This has involved processing large open source documents pertaining to the 9/11 attacks, including the publicly available 9/11 commission report. The goal is to permit interactive text mining on this corpus leading to intelligence from open sources. We have focused on two types of text mining queries: (i) concept chains (i.e., looks for best paths) and (ii) scenario detection (looks for patterns). A concept chain query on this corpus looks for best paths between two entities (e.g., connecting the trucking industry and foreign banks). This reveals various paths crossing multiple documents (e.g., a truck parts manufacturer through an insurance claim to a foreign bank). A scenario query looks for patterns of activities that can lead to a security scenario. For example, the mention of a person's interest in a chess club, the announcement of a small business formation and an upcoming lecture at the public library may seem unrelated at first. However, the knowledge that (i) the chess club meets at a public library, (ii) the chess enthusiast is an employee of the newly formed business, (iii) the business owner is a native of country X which engages in state-sponsored terrorism, and (iv) the library has scheduled a controversial speaker critical of country X. This pattern of activity may warrant closer inspection.

The initial focus has been on developing infrastructure (e.g., testbeds for validation and verification, domain ontologies, and especially building the CCG).

UIR Toolkit:

- We have tuned an information extraction engine to detect salient concepts automatically in documents and map them into an ontology. The concepts are detected in a data-driven process, i.e. they are not predefined. However the ontology nodes are predefined, and therefore automatic techniques have been developed to perform this mapping.
- We have demonstrated the feasibility of constructing a new type of index for information retrieval for a modest-size collection (e.g., 2,000 documents). Our CCG integrates the traditional bag-of-words model with a higher-level conceptual model. We have shown this includes the information retrieval model and can also perform certain retrievals that a traditional information retrieval system cannot.

UIR Prototype:

Finally, we have developed a UIR prototype with a Graphical User Interface (GUI), etc. Users can retrieve documents based on concepts (categorized into entity types), retrieve document sets based on concept chains, see evidence trails connecting concepts, and see graphical descriptions of the neighborhood surrounding concepts. The specific functionality includes:

- We have demonstrated concept-based search, i.e. search using ontology terms rather than keywords.
- We have demonstrated concept browsing: a visual representation of the concepts closely associated with some scenarios.
- We have demonstrated concept chain retrieval. This exposes links between concepts crossing multiple documents. It shows connecting evidence trails.
- We have demonstrated concept chain searches. E.g. find the best match connecting Taliban, the USS Cole, and San Diego.

The next stage of our development focuses on:

- Improving the concept chain generation process by incorporating additional weights reflecting the novelty of a concept, as well as more sophisticated mathematical models.
- Adapting the software to the FAA domain: this involves porting the ontology, as well as processing new content
- Defining certain scenarios known to be of interest to the FAA: implementing graph mining techniques to find instances of these
- Initial work on detecting scenarios of interest in a data-driven manner

Related Work

> Text mining focuses on discovering new knowledge such as trends and patterns buried in a huge collection of text documents. This is in contrast to typical data mining that deals with finding patterns and associations in structured data. Common criticisms of data mining tools are the tools: (i) provide an overwhelming amount of spurious data and (ii) reveal information that is already known. The goal of our research is to provide more focused text mining capabilities that allow the user to direct the query in a meaningful way.

Two domains that have seen significant activity in text mining include the biomedical domain and the intelligence analysis domain. An example of the intelligence analysis domain is the DARPA Evidence Extraction and Link Discovery (EELD) ~~http://www.darpa.mil/iao/EELD.htm~~ program. From an historical perspective, the work of Swanson and Smalheiser [Swanson-1988, Smalheiser-1996] is relevant. They proposed that combining existing, though unconnected, bibliographic information results in new knowledge. One publication may present a relationship between A and B while another reports a relationship between B and C. If no one has reported on the association between A and C, this can be considered a new finding of scientific interest. The most important notion in this approach is that there exists a hidden connection (e.g., B) between these two pieces (e.g., A-C) of information.

Early text mining efforts were mostly based on information retrieval techniques but have already produced interesting results in: (i) Irregular Behavior Detection, including fraud detection; (ii) Mining Associations and Trends Analysis; (iii) Detecting Potential Missing Links; and (iv) Novelty Detection. [Berry-2004] gives a good survey of these information retrieval techniques used in text mining, including latent semantic indexing. Latent semantic indexing is used to model document content as a feature vector comprised of weighted “concept” vectors; this in turn, permits more conceptual query matching as opposed to keyword matching. Latent semantic indexing features (from textual documents) have been combined with features obtained from structured data in tasks such as

predictive modeling. It suffices to say that latent semantic indexing can be viewed as a form of conceptual feature extraction for text. While these techniques are useful in text mining and trend analysis, they are limited to doing this at the document level. For other text mining tasks, such as finding the most plausible connection between two organizations, finer granularity is required.

There has been work on discovering connections between concepts across documents using social network graphs, where nodes represent documents, and links represent connections (typically URL links) between documents. However much of the work on *social network analysis* has focused on different types of problems, such as detecting communities [Gibson-1998]. [Faloutsos-2004] is the work that is closest to the research presented here, at least in its goals. The authors model the problem of detecting associations between people as finding a connection subgraph and present a solution based on electricity analogues. However there are several differences that should be noted. The most notable difference is the reliance on URL links to establish connections between documents.

Our approach extracts associations based on content (textual) analysis. Second, the connection subgraph approach presents all paths together, while our approach presents the paths individually. This allows greater user input in determining the *best* paths, including recency, novelty, semantic coherence, etc. Third, the approach presented here attempts to generate an explanation of the chains, whereas the connection subgraph approach does not. Finally, the connection subgraph solution only addresses named entities whereas this approach extends to general concepts.

More recently, the DARPA Evidence Extraction and Link Discovery program has resulted in text mining efforts that use more sophisticated information extraction output such as named entities, relationships and events corresponding involving key entities [Weiss 2004]. Such systems typically use information extraction tools to extract salient entities and relationships; these are then input to either visualization or link analysis tools. Examples of the latter include VisuaLinks software (www.visualanalytics.com). This often leads to more spurious patterns than significant ones. Recently, there have been attempts to guide the mining process by using models, reflecting patterns of interest. Such models require the underlying data to be well structured and noise-free, which is not the case in data obtained from information extraction systems. Probabilistic Relational Models and more sophisticated variants [Heckerman 2004] have been proposed as a technique for combining relational models with Bayesian networks [Heckerman 1999]. While these techniques could potentially discover new trails of information corresponding to a specific entity, they are limited in that the trails are all *instance-based*. In other words, one can track activities pertaining to specific individuals, organizations, or locations. It is difficult to use these tools to find trails connecting, or going through generic concepts. For example, consider the following text snippets coming from three separate documents: (i) *the pilot was a private pilot with single and multi-engine land ratings*, (ii) *unable to fully extend one main landing gear, the private pilot retracted the other gear, …*, and (iii) *other gears in the accessory case exhibited unusual wear patterns*. Based on these snippets, one may be able to derive a plausible correlation between private pilots and accidents due to misuse of landing gear. It should be noted that the private pilot referred to in the first document is not the same one mentioned in the second document. However, they could have been the same, and it is this type of plausible correlations that text mining needs to expose.

This leads us to the novelty of the UIR solution presented here. First, it defines a new content representation that is able to: (i) store information about both instances and general concepts such as *pilot*, and (ii) explicitly store connections between concepts across documents. It is the latter that leads to interesting trails of information.

Publications/Briefings:

Initial work in the biomedical domain was published at an Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) workshop on biomedical text mining. A paper describing the application of this work to the intelligence community will be presented at a workshop on Context in Information Retrieval, in conjunction with the Context-2005 conference being held in France later this summer. Currently, conference papers for submission to the Association for Computing Machinery's Conference on Information and Knowledge Management (CIKM 2005) as well as the Journal of Intelligence Community Research and Development (JICRD) and ACM Transactions on Information Systems (TOIS) journals are being prepared. The Principal Investigator has briefed the FAA regarding the project and demonstrated the prototype system; this received very positive feedback.

Project Website:

www.cedar.buffalo.edu/~rohini/UIR

Funding History:

This project has been funded both by the FAA and the National Science Foundation (NSF) since the summer of 2003. The NSF funding focuses on fundamental research and intelligence community applications. The FAA funding developed the UIR prototype as it applies to FAA data.